Lecture 13 — October 6, 2016

Prof. Eric Price *Scribe: Kiyeon Jeon and Loc Hoang*

1 Overview

In the last lecture we covered the lower bound for p^{th} moment ($p > 2$) and the concepts of packing numbers, covering numbers, and metric entropy.

In this lecture we discuss **Maurey's Empirical Method** for covering numbers and begin moving into compressed sensing by starting with **Restricted Isometric Property (RIP) matrices**.

2 Covering Numbers and Maurey's Empirical Method

2.1 Introduction

Last lecture, we discussed the problem of getting a covering number N for L_1 balls using L_2 balls.

$$N(\epsilon, B_1^d, ||\cdot||_2) \tag{1}$$

Using a volume argument, we were able to establish the following result.

$$N(\epsilon, B_1^d, ||\cdot||_2) \leq N(\epsilon, B_1^d, ||\cdot||_1) \tag{2}$$

$$N(\epsilon, B_1^d, ||\cdot||_1) \leq (1 + \frac{2}{\epsilon})^d \tag{3}$$

The first inequality comes from the fact that L_1 balls take up less space than L_2 balls, so it would take more of them to do the covering.

From this, we could get metric entropy $\log N$.

$$\log N \geq d \log \frac{1}{\epsilon} - \frac{d}{2} \log d \tag{4}$$

We can then deduce that $\log N = \theta(d \log \frac{1}{\epsilon})$ if $\epsilon < d^{-\frac{1}{2}-\Omega(1)}$. We are interested in the case where $\epsilon > \frac{1}{\sqrt{d}}$, and this will be examined in the next section.

2.2 Using Maurey's Empirical Method for Covering Numbers

Using Maurey's empirical method, we will show the following:

Theorem 1. *When* $\epsilon > \frac{1}{\sqrt{d}}$, $N \leq (2d+1)^{O(1/\epsilon^2)}$

As a result, $\log N \lesssim \frac{1}{\epsilon^2} \log(d)$.

Proof. Let's cover the following set:

$$B_1^{d,+} = \{x \in \mathcal{R}^d \mid \|x\|_1 \leq 1 \text{ and } x_i \geq 0 \; \forall i\}$$

The above set means that $\sum x_i \leq 1 \; \forall x_i \geq 0$.

We can think about a probability distribution over $\{e_1, \ldots, e_d, 0\}$:

$$z = \sum_{i=1}^{d} x_i e_i + (1 - \|x\|_1) \cdot 0$$

This implies the following probabilities.

$$\mathbb{P}[z = e_j] = x_j \; \forall j \in [d]$$
$$\mathbb{P}[z = 0] = 1 - \|x\|_1$$

With these, we can get a mean of the probability distribution.

$$\mathbb{E}[z] = \sum \mathbb{P}[z = e_j] \cdot e_j + \mathbb{P}[z = 0] \cdot 0 = \sum x_j \cdot e_j = x$$

We will draw t samples $z_1, \ldots, z_t$ from the distribution where each z is some e_i. After drawing the samples, we can take the average of the samples:

$$\bar{z} = \frac{1}{t} \sum_{i=1}^{t} z_i$$

We want to show that $\mathbb{E}[\|\bar{z} - x\|_2^2] \leq \epsilon^2$. If we can do this, then if we take all possible $\bar{z}$, we get an ϵ-cover of the space using those $\bar{z}$ since then all x we can choose will be within ϵ of some point in the cover by what we argue above.

We work this out below:

$$\begin{aligned}
\mathbb{E}[\|\bar{z}-x\|_2^2] &= \mathbb{E}[\sum_{j=1}^{d}(\bar{z}_j - x_j)^2] \\
&= \sum_{j=1}^{d} \text{Var }(\bar{z}_j)(\text{definition of variance}) \\
&= \sum_{j=1}^{d} \frac{1}{t^2} \text{Var }(\sum_{i=1}^{t}(z_i)_j)(\text{independent variances means you can sum them}) \\
&= \frac{1}{t}\sum_{j=1}^{d} \text{Var }((z_1)_j) \\
&= \frac{\sum_{j=1}^{d} x_j}{t} = \frac{\|x\|_1}{t} \\
&\leq \frac{1}{t}(\text{by our original definition/choice of } x)
\end{aligned}$$

Now, let $t = 1/\epsilon^2$. We have the following desired bound:

$$\mathbb{E}[\|\bar{z}-x\|_2^2] \leq \epsilon^2$$

This implies that there exists $\bar{z}$ with $\|\bar{z}-x\|_2 \leq \epsilon$ if $t = 1/\epsilon^2$. Then we can pick all $\bar{z}$ to create our ϵ-cover. The number of such $\bar{z}$ is $\leq (d+1)^t = (d+1)^{O(1/\epsilon^2)}$ (sample t zs, and there are $d+1$ choices for each z, then take the mean).

Therefore, we have a bound on packing number and a bound on the metric entropy.

$$N \leq (d+1)^{O(1/\epsilon^2)}$$
$$\log N \leq \frac{1}{\epsilon^2}\log(d+1)$$

This implies $\log N \lesssim \frac{1}{\epsilon^2}\log(d)$ as desired.

□

Note that this could be extended to cover a larger space (note we only cover $B_1^{d,+}$, which is a positive space). The basic proof idea will still go through if we decided to extend it to larger cases.

3 Restricted Isometric Property (RIP) Matrices

We move the discussion towards RIP matricies, which will move us closer to compressed sensing.

3.1 Definition of RIP

We say a vector $x \in \mathcal{R}^n$ is k-sparse if $|\text{supp}(x)| = \|x\|_0 \leq k$.

Define $T_k := \{x \in \mathcal{R}^n | \|x\|_0 \leq k, \|x\|_2 \leq 1\} \subseteq \mathcal{R}^n$, or the set of all vectors that are k-sparse as well as have an L_2 norm that is less than 1. We want to determine the a bound on the metric entropy of T_k, $\log N(\epsilon, T_k, \|\cdot\|_2)$. To do so, we look at $\log N(\epsilon, T_n, \|\cdot\|_2)$ (T_n ignores sparsity). From what we covered last class, we can determine that $\log N(\epsilon, T_n, \|\cdot\|_2) \leq (1 + \frac{2}{\epsilon})^d$. From here, we can take a union bound over k-dimensional subspaces (by using $\binom{n}{k}$) to bound our original packing number:

$$\begin{aligned} N(\epsilon, T_k, \|\cdot\|_2) &\leq \binom{n}{k}(1 + \frac{2}{\epsilon})^k \\ &\leq (\frac{en}{k})^k(1 + \frac{2}{\epsilon})^k \\ \log N(\epsilon, T_k, \|\cdot\|_2) &\leq k \log(\frac{n * 2e}{\epsilon k}) \end{aligned}$$

This bound is good as it is something that depends reasonably on k. Using this, we can get a bound on RIP matrices.

Definition 2. *A matrix $A \in \mathcal{R}^{m \times n}$ is a (k, ϵ) RIP (Restricted Isometry Property) matrix of order (k, ϵ) if $\forall k$-sparse x, $\|Ax\|_2 = (1 \pm \epsilon)\|x\|_x$*

RIP matrices are useful for recovery of vectors:

Theorem 3. *Let $y = Ax + e$ where x is k-sparse and A has $(O(k), .1)$-RIP. Then, one can recover an $\hat{x}$ such that $\|\hat{x} - x\|_2 \leq O(\|e\|_2)$*

Also, let $y = Ax$ where x is not k-sparse and A has RIP. Then, $\|\hat{x} - x\|_1 \leq O(\|x - x_k\|_1)$

We can compare the first bound on $\hat{x}$ to Count-Sketch. In Count-Sketch, we have Ax where x is not k-sparse. Then, w.h.p. we get get $\hat{x}$ with $\|\hat{x} - x\|_2 \leq (1 + \epsilon)\|x - x_k\|_2$. The second bound of the theorem is comparable to Count-Min.

Basically, RIP lets us observe good results once we have selected a good A with the RIP.

3.2 Analysis of Number of Rows (m) for a RIP Matrix

How large does m need to be? We need to determine how many rows m there needs to be in a matrix A in order to satisfy (k, ϵ) RIP.

Pick A as an i.i.d. (sub)gaussian matrix. For any fixed x, we have that $\|Ax\|_2^2 = (1 \pm \epsilon)\|x\|_2^2$ w.p. $1 - \exp^{-\epsilon^2 m \cdot \Omega(1)}$.

We want to be able to get a bound for **all** x, not fixed x. We could use a union bound, but the problem is that there are infinitely many x. This is where we can use metric entropy: close vectors x_i (in some space) will be similar in behavior. Therefore, it suffices to take a union bound over an ϵ-cover of T_k. $m = O(\frac{1}{\epsilon^2} k \log \frac{n}{\epsilon k})$ will work. This m is needed so that we can use the Johnson-Lindenstrauss bound on the L_2 norm later in the proof below.

Proof of Existence of an A satisifying (k, ϵ) RIP Let C be an ϵ-cover of T_k built as $\binom{n}{k}$ ϵ-covers of subspaces with support size k. Now, take $x^* \in T_k$.

$$x^* = x_1 + x', \|x'\| \leq \epsilon, x_1 \in C$$

From the fact that $x_1 \in C$ and because of a special property of our cover C (the way it was built), we have that $|\text{supp}(x_1) \cup \text{supp}(x^*)| \leq k$. From this we have that $\|x'\|_0 \leq k$. since we are "rounding" within the same subspace. Now that $\|x'\| \leq \epsilon$ (since x^* is at most ϵ far from x_1 since C is an ϵ-cover) and $\|x'\|_0 \leq k$, we can say $x' \in \epsilon T_k$.

Similarly to x^* above, we can find $x_2 \in C$ and $\|x''\| \leq \epsilon$ satyisfying

$$x' = \epsilon(x_2 + x'')$$

Again, with a similar argument to the one we made for x' above, we can obtain $\|x''\|_0 \leq k$. Again, we can find an $x_3 \in C$ and $\|x'''\| \leq \epsilon$ like above, and we can do this over and over.

As we do this process continuously, we derive the following:

$$\begin{aligned}
x^* &= x_1 + \epsilon x_2 + \epsilon^2 x_3 + \cdots \epsilon^{l-1} x_{l-1} + x^{(l)} \\
&\text{where } \|x^{(l)}\| \leq \epsilon^l, \|x^{(l)}\|_0 \leq k, x_i \in C \ \forall i = 1, \ldots, l-1
\end{aligned}$$

Now, let $m = O(\frac{1}{\epsilon^2} k \log \frac{n}{\epsilon k})$. This size of m allows us to use Johnson-Lindenstrauus for a tail bound, specifically $\|Ax\| = (1 \pm \epsilon)\|x\| \ \forall x \in C$ by a union bound (see lecture 10 of the Fall 2014 version of the class). We use this below.

$$\begin{aligned}
\|Ax^*\| &\leq \|Ax_1\| + \epsilon\|Ax_2\| + \epsilon^2\|Ax_3\| + \cdots \\
&\leq (1+\epsilon)(1 + \epsilon + \epsilon^2 + \cdots) \\
&\leq \frac{1+\epsilon}{1-\epsilon} = 1 + O(\epsilon) \\
\|Ax^*\| &\geq \|x^*\| - O(\epsilon)
\end{aligned}$$

So, $\forall x \in T_k$,

$$\begin{aligned}
&\|x\| - O(\epsilon) \leq \|Ax\| \leq (1 + O(\epsilon)) \\
&\forall x \in T_k \text{ with } \|x\| = 1, \|Ax\| = (1 + O(\epsilon))\|x\| \rightarrow \\
&\forall x \in T_k \ \|Ax\| = (1 + O(\epsilon))\|x\| \rightarrow \\
&A \text{ satisfies } (k, O(\epsilon))\text{-RIP if } m = O(\frac{1}{\epsilon^2} k \log(\frac{n}{\epsilon k}))
\end{aligned}$$

3.3 Further Considerations and Extensions

3.3.1 Other Considerations for A

1. What about small integer entries in A?

   What if $A_{ij} \in \{\pm\frac{1}{\sqrt{m}}\}$ i.i.d.?

   It turns out the proof will work for any subgaussian generated matrix. We only used the Gaussian generated matrix for the JL property for the tail bound.

2. Can we get A to be sparse so we can do computations faster?

   This can't happen: shown on homework.

3. Can we still hope to store A more efficiently with limited independence?

   Probably not, but note that A (a matrix with the RIP property) can be a fixed matrix (i.e. find once, then reuse).

4. Can we get an explicit matrix?

   We can show a $k^{1.9\times}$ construction.

5. Are there other matrices that fast to store and compute?

   We look at Fourier matrices in the following section.

3.3.2 Extensions

Question: How might we use the Fourier matrix to get a RIP matrix that is easier to deal with?

We have $\|Fx\|_2 = \|x\|_2$ (note it preserves the L_2 norm) where $F_{ij} = \frac{1}{\sqrt{n}} \exp^{2\pi\sqrt{-1}\frac{i\cdot j}{n}}$. By choosing $\Omega \subseteq [n]$ at random (i.e. choose random rows), we get $F_\Omega \times \sqrt{\frac{n}{m}}$. If $|\Omega| \geq \frac{1}{\epsilon^2} k \log n \log^2 k$, then F_Ω has (k, ϵ)-RIP with "good" probability.

(note that there have been better results for the last term in the inequality above: $\log^5 k['06], \log^3 k['08], \log^2 k['16]$)

	Gaussian	Fourier
space to store	mn	m
time to multiply	mn	$n \log n$
time to multiply by k-sparse vector	mk	$\min(mk, n \log n)$

We can also use Circulant matrices. Such a matrix C_v is shown below.

$$\begin{array}{ccccc} v_1 & v_2 & \cdots & v_{n-1} & v_n \\ v_2 & v_3 & \cdots & v_n & v_1 \\ \vdots & \vdots & & \vdots & \vdots \\ v_n & v_1 & \cdots & v_{n-2} & v_{n-1} \end{array}$$

Similar to what we did for Fourier matrices, we choose $(C_v)_\Omega$ where Ω is **arbitrary** on Circulant matrices. Unlike Fourier, where row selection had to be random, we can choose whichever rows we want without the need for randomess. v, however, is random (e.g. Gaussian).

The number of rows m is $O(k \log n \log^3(k \log n))$.